DOCUMENT RESUME

ED 302 570 TM 012 574

AUTHOR Sinkavich, Frank J.
TITLE Multiple Choice Examination Performance and Metamemory Accuracy in Good and Poor Students.
PUB DATE 88
NOTE 29p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cognitive Processes; *Confidence Testing; Educational Psychology; Graduate Students; Higher Education; Memory; *Metacognition; Multiple Choice Tests; Self Evaluation (Individuals); Test Format

ABSTRACT

The relationship between metamemorial accuracy and student test performance was assessed in a graduate educational psychology class. Metamemory is defined as information about the content of one's own memory. Adult students in two courses (N=67) were asked to select multiple choice answers in two midterms and a final examination and to rate their confidence in the answers on a five-point Likert scale. At the end of the test, students were asked for a prediction of the percentage of items they thought they answered correctly. They were then given ten replacement items and were allowed to replace any ten test items with any of these replacements. Those who had a high mean degree of confidence in their answers had higher examination scores than those with low confidence in their answers. There was a significant difference between good and poor scorers in their ability to predict what they did and did not know. Good students, as defined by scores on the final examination, appeared to have better metamemorial accuracy than the poor students. The use of replacement items allowed students to improve their scores significantly. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302570

Multiple Choice Examination Performance and Metamemory Accuracy in Good and Poor Students

Frank J. Sinkavich

University of Pittsburgh

The purpose of this study is to assess the relationship between metamemorial accuracy and student test performance in a graduate educational psychology class. This study examined the accuracy of metamemory in adult students by assessing their degree of confidence in their answers to multiple choice test items and its relationship to their actual test item performance. This study is concerned with knowledge of one's acquiring of psychology in a graduate course. Students from two graduate educational psychology courses took two mid-terms and one final examination. There were 38 students in the winter term course and 29 students in the spring session course. The students were not psychology majors.

On each exam, for each item, students were asked to select their multiple choice answer and then to rate their confidence in that answer on a five-point Likert scale. After their concurrent confidence rating, at the end of the exam, on the final page, students were asked to make an overall and retrospective prediction of the percent of items they thought they got correct on the test. The procedure for the final exam had one additional step. Students were asked to complete their exam, both answers and confidence ratings, and they were then given a sheet of ten replacement items. They could replace any ten items on the final exam with any of the ten replacement items.

Results indicate that students who demonstrated a high mean degree of confidence in their answers have higher examination scores when compared to those students who expressed a low mean degree of confidence in their answers. Results also indicate that a significant difference exists between "good" and "poor " students in their ability to predict what they know and do not know. "Good" students were better able to predict their item performance when compared to "poor" students. This result implies that "good" students have better metamemorial accuracy than do "poor" students on multiple choice examinations. Finally, results indicate that through the use of replacement items students can significantly improve their scores on a test, or their standing in the class.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

FRANK J. SINKAVICH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012574

Introduction

The question of what people know about their own thought processes is an old one in the field of psychology (Bartlett, 1932; Jenkins, 1973; Neisser, 1967,1976; Norman, 1973). In recent years, the area of metacognition has come under increased investigation by cognitive, developmental, and educational psychologists. Metacognition is defined as knowledge of one's own cognitive processes (i.e., knowing what one knows) and state of knowledge. The assumption is that metacognition underlies the ability of persons to monitor their current levels of understanding and learning.

Most of the research on metacognition has been centered around children of elementary school age (Brown, Bransford, Ferrara, & Campione, 1983; Brown, 1978; Ghatala, 1986). Research has been focussed on elementary school students because this is the age when memory and intelligence are developmental issues (Piaget and Inhelder, 1973) and start to become the center of attention for educators. Baker and Brown (1984) discuss the relationship between metacognitive skills and reading comprehension with elementary students.

Except for the gerontological work, and the life-span studies (see Lachman and Lachman, 1978), most studies of what we know about our own knowledge (i.e., metacognition) have assessed the metacognitive capabilities of children or exceptional populations. Studies have either used experimentally learned materials (e.g.,. nonsense syllables), or to a lesser degree they have been concerned with knowledge acquired in the course of day-to-day living (e.g., world knowledge). No studies to date have been concerned with students knowledge of what they have learned from a course of formal instruction. To answer test items that measure knowledge acquired from formal instruction requires retrieval, inference, and metamemorial control processes (Lachman and Lachman, 1978).

Research on effective learners, and good and poor readers, writers, and problem solvers has shown skilled individuals and good achievers are the better "metacognizers" (Brown, Bransford, Ferrara, & Campione, 1983; Brown, 1978; Ghatala, 1986; Baker & Brown,1984; Flavell,1976 & 1979; Bransford, 1979) That is, these persons are better able to self-regulate their own learning, which in the present study's context (taking a test) would predict that "good " students have higher metamemory accuracy than "poor" students.

The concern of the present study was with knowledge and information that is relatively permanent (learning), nonepisodic, and has not been studied for the sole purpose of demonstrating it can be remembered, but

rather has been acquired during a semester of formal classroom instruction and day-to-day experience. This study is concerned with the degree to which students know what they do and do not know. The central assumption is that metacognition and the processes of metamemorial control direct both retrieval and inferential activity and the resulting behavioral actions. The resulting actions studied here are the answers given in a multiple choice examination.

Statement of the problem

The purpose of this study was to assess the relationship between metamemorial accuracy and student test performance in a graduate educational psychology class. This study examined the accuracy of metamemory in adult students by assessing their degree of confidence in their answers to multiple choice test items and its relationship to their actual test item performance. The underlying assumption is that metamemory can be inferred from the accuracy of these predictions (Lachman and Lachman, 1978; Slif, Weiss, and Bell, 1985; Hart 1965).

This study will investigate the relationship between two variables: (a) Metamemory; as measured by a student's predictions of success on a classroom examination; and (b) actual performance on the examination.

Hypotheses

1. Students who express a high mean degree of confidence in multiple choice answers have higher examination scores than students who express a low mean degree of confidence.

2. Students who achieve high scores on classroom examinations("good" students) are better predictors of what they do or do not know on a test than are low achieving students ("poor" students).

3. Good students use metamemory judgements to make correct decisions about item responses they do or do not know when permitted to choose replacement items on an examination whereas poor students do not.

Metacognition

Metacognition refers to knowledge about cognitions and the executive decision making process in which the individual must both carry out cognitive operations and oversee his or her progress (Meichenbaum, 1984). Ann Brown (1978) refers to metacognition as "knowing about knowing"; the subjects' awareness of their own cognitive processes and how these processes work. In other words, metacognition refers to one's *knowledge* and *control* of cognitive processes.

The activities that this study defines as metacognitive are those found in the developmental literature and consists of those processes used to regulate and oversee learning and performance. These processes include planning activities prior to starting a problem (e.g., predicting outcomes, scheduling strategies), monitoring activities during the learning process (e.g., testing, revising, and re-scheduling strategies), and checking outcomes (e.g., evaluating the outcome of any strategic actions against criteria for efficiency and effectiveness; Brown, et al., 1983; Brown, 1978).

Metamemory

According to Lachman and Lachman (1978), metamemory is a collection of memory control capacities, including the assessment of whether a particular piece of knowledge is available, how likely one is to access it, and how certain one is that it is accurate. In other words, metamemory is knowledge about one's own information store (i.e., the content of memory). Metamemory is a narrower term than metacognition.

Prediction of success

Confidence in item correctness is measured by a five-point Likert-type scale (Appendix B) with ratings from -2 to +2. Students were instructed to indicate on the scale how *correct they feel* their answer is.

Academic performance

Performance is defined as scores on the classroom examinations given during the course and on the initial examination given on the first day of instruction to assess the student's knowledge of educational psychology prior to instruction.

Feedback

Feedback will be defined as the instructor's report to the students after each examination on how well they predicted their success on each individual test item and on the overall outcome of the examination. This report will be a microcomputer generated document (Appendix A) containing such data as (a) student name; (b) exam score; (c) number correct; (d) number incorrect; (e) incorrect hits; (f) correct hits; (g) prediction mean; (h) prediction outcome; (j) actual outcome; and (k) a graphic display (histogram) of the prediction data.

Summary

Optimal use of previously acquired knowledge is in part a function of metamemorial accuracy. The optimal use of such knowledge is important to the field of education. Knowledge of course information is in itself, of course, insufficient; students must be able to use it. A significant component of being able to use knowledge optimally is accurate judging of whether or not you possess it. This study examines the relationship between judgements of "knowing" and actual test item performance.

Review of the Literature

Feeling of knowing

Knowledge about one's own information store (i.e., what is in one's memory) is mirrored by a phenomenon called "feeling of knowing" (Lachman & Lachman, 1978). According to Hart (1965), the feeling-of-knowing (FOK) experience is an accurate indicator of the content of one's memory. Hart (1965) discusses a process he calls memory-monitoring (i.e., metamemory). He defines this process as referring to the intervening process that produces the feeling-of-knowing and feeling-of-not-knowing judgements that one has when trying to answer questions. This memory-monitoring or metamemorial process appears to be an important one, contributing significantly to the efficiency of the human information-processing system.

People are not completely blank, even when they are not able to answer a difficult question. Most of the time they have definite feelings about their knowledge. That is, they have feelings about what they know or don't know. According to Hart (1965) feelings of knowing can sometimes be very strong; a person will feel that an elusive memory is close (i.e., right on the tip-of-the-tongue). Tip-of-the-tongue experiences and feelings of knowing of lesser intensities are common, occurring in every day life with many types of memory material. For example, If you ask a person "Who was the Secretary of State in the Kennedy administration?" they will respond, "let me think, I know but...". If you ask this same person could they recognize the name of the Secretary of State under the Kennedy administration if shown a list of names, they would reply "yes".

Tip-of-the-tongue experience has been recognized for many years by psychologists. Woodworth and Schlosberg (1954) summarized the early research and investigations into the phenomenon. Hart (1965) states that the early investigations were limited in several respects: (a) Only intense tip-of-the-tongue experiences were studied, not the more general and ubiquitous feeling-of-knowing experiences; (b) these investigations were nonquantitative and unsystematic collections of instances; (c) the investigations did not address the most important question regarding tip-of-the-tongue of feeling-of-knowing experiences, that is, are they accurate? The early investigators took the phenomenon as given and attempted to study how subjects retrieved or searched for information they did not have but felt they knew. Hart (1965) states that this departs from the study of FOK phenomenon itself. Asking how

FOK memories are retrieved presupposes that the FOK is an accurate indicator of what is in memory. Hart's research tests this presupposition. It is for this reason (FOK is an accurate indicator of memory content) that the proposed study of metamemorial accuracy is grounded in the pioneering work of Hart.

Feeling of knowing research

Accuracy of feeling-of-knowing can be assessed by comparing the proportion correct on FOK items with the proportion correct on feeling-of-not-knowing items for each student on a given examination (Hart, 1965). To answer the question about FOK experiences Hart (1965) used a research paradigm that allowed for the experiences to be produced and their accuracy evaluated. He made use of one of the best established facts of verbal learning: "recognition exceeds recall." People can almost always recognize more answers than they can produce. From this fact Hart applied the following recall-judgement-recognition (RJR) paradigm as a way of studying the accuracy of FOK experiences: (a) give the subjects a test of recall and, for those items that they cannot answer, instruct them to make a judgement about whether or not they feel they know the correct answer well enough to recognize it among several wrong alternatives; (b) then give the subjects a multiple-choice recognition test covering the same items that appeared in the test of recall.

According to Hart (1965) if the FOK experience is an accurate indicator of memory storage, the subjects should do better on those recognition-test items which they know but cannot recall than on those items which they feel they do not know. The accuracy was assessed by comparing the proportion correct on feeling-of-knowing items with the proportion correct on feeling-of-not-knowing items for each subject. Hart (1965) conducted two experiments employing the same RJR two step procedure. The only difference between experiment 1 and experiment 2 was the number of questions used (50 and 75 respectively) and the kind of FOK judgements obtained from the subjects, (a) dichotomous (i.e., yes or no) and (b) graded (a six point rating scale) respectively.

The results of this research by Hart (1965) indicated that feeling-of-knowing judgements can be viewed as relatively accurate indicators of the information that is and is not in memory (proportions for items correct, according to FOK ratings, were significantly different from the chance level. The range of the proportion correct was .48 -.78).

In other words, Hart has answered the question: Are FOK experiences accurate? His results in the 1965 study indicate that they are.

Hart (1965) states that many of his multiple-choice alternatives (recognition test) did not seem equally likely. In other words, a person ignorant of the correct answer but possessing some related knowledge of the field covered by the question might be able to narrow the alternatives down to two or three. This would result in raising the guessing probability to .50 or .33 (.25 would be a guessing probability if all four alternatives were equally likely). For example a sample question (Hart,1965) "Who wrote *The Tempest*"? , many subjects might be able to narrow their guesses to the English writers, Jonson and Shakespear from the four possible answers: Moliere, Strindberg, Jonson, Shakespear. This increase in guessing probability would result in an increase in the percent correct of FOK items.

Students predictions of what they can and cannot remember are not perfect and there are individual differences among students (e.g. experience) that can effect accuracy. Even given this, according to Hart (1965), the feeling-of-knowing phenomenon appears to be powerful and reliable. When students feel definitely that they do not know an answer to a question, they score on those items approximately at chance recognition level (e.g., .25 the guessing probability of four equally likely multiple-choice answers). Conversely, when they feel that they do know the answer to an item, their score is three times the chance level (Hart, 1965). Hart (1966), addresses some methodological problems with the RJR paradigm (recall-judgement-recognition) on which his assessment of feeling-of-knowing is based. He states that overcautious withholding of answers by students constitutes a significant source of feeling-of-knowing inflation. Hart recommends that all research using the RJR paradigm should instruct students to provide answers to every question, even if a guess is required.

In summary, the work of Hart has shown that feeling-of-knowing judgements can be viewed as relatively accurate indicators of the information that is and is not in one's memory . Hart's work indicates that the feeling-of-knowing experience is an accurate (better than chance probability) indicator of memory storage.

Metamemorial control processes

According to Lachman and Lachman (1978), metamemorial control processes are the hypothetical mechanisms that are responsible for directing one's efforts to retrieve information and draw inferences. The feeling-of-knowing phenomenon can be compared to metamemorial control processes in that the FOK experience could be considered a control process that allows a person to identify whether or not they possess knowledge of particular information or experiences. Metamemorial control processes are assumed to be a part of an individual's willingness to act on retrieved or constructed information. These processes are employed in part to ascertain whether sufficient related information is stored in memory to warrant an inference.

Actualization of knowledge

Actualization of course knowledge requires a complex interaction of recall, retrieval, inference, and the use of non-course or outside course experience in a person's attempt to answer an examination item, whether the item is a knowledge item or an application item.

Figure 1 shows the relationship between metamemorial control processes and the three subsystems implicated in the actualization of course knowledge. This hypothesized organizational framework links together recall, inference, and experience. In this organizational framework, all three subsystems are hypothesized to be under metamemorial control. Each of the three subsystems (recall, experience, and inference) are linked together at one level (dotted lines) where they can interact. At a second level (solid lines), each subsystem is linked to the metamemorial control processes. At this level each can act alone or in concert with one another under the direction of the metamemorial control processes to aid the student in determining what he or she does or does not know.

An organizational framework for exploring questions about metamemory accuracy.
(Adapted from Jenkins, 1978; and Bransford, 1979)

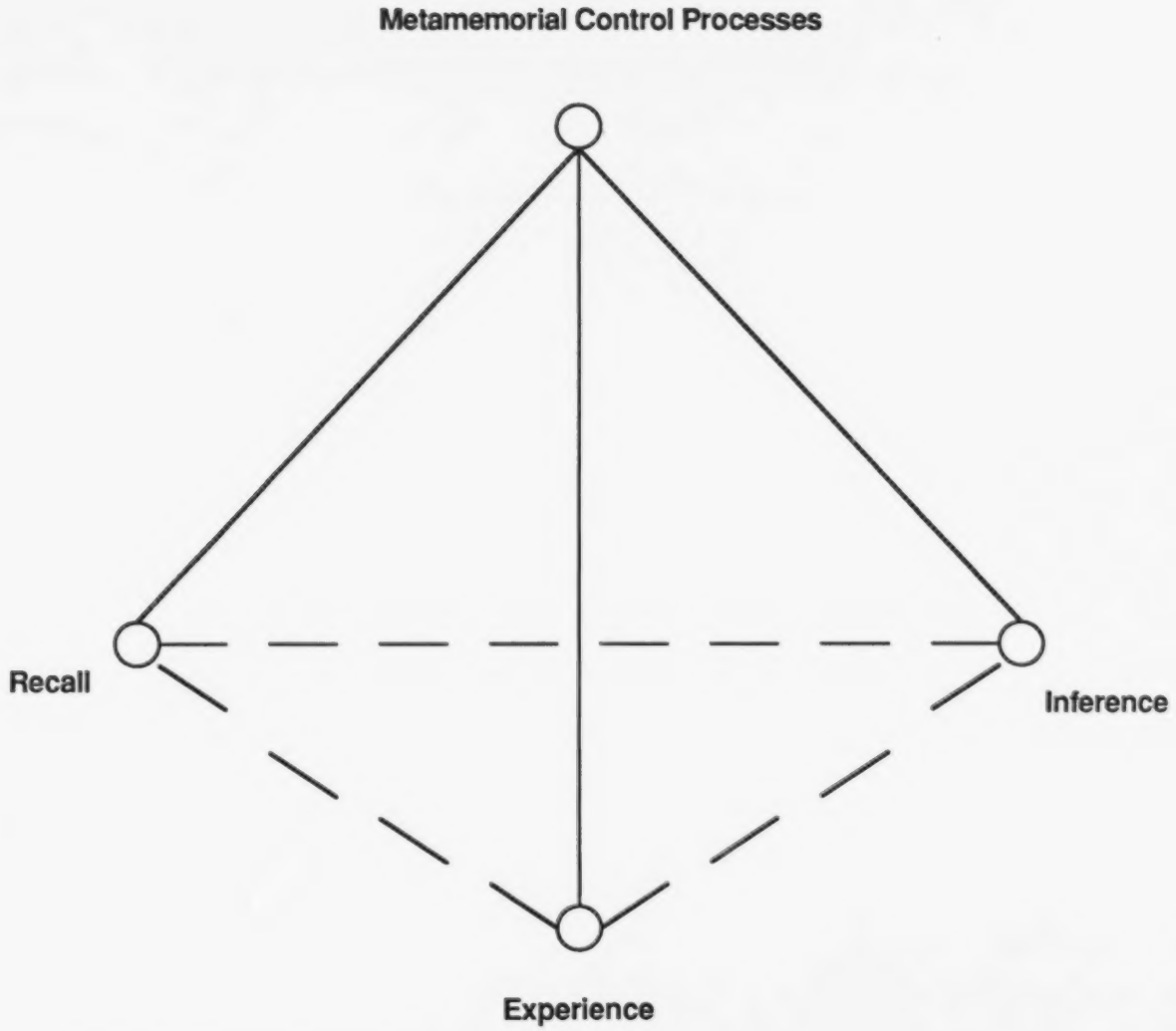

Figure 1 Hypothetical relationship among three subsystems implicated in the actualization of course knowledge

Metacognition, metamemory, and test item performance

Metacognitive processes (i.e., planning, monitoring, & checking outcomes) are central to learning and to performance. Planning includes those activities engaged in prior to problem solving (e.g. scheduling activities and predicting outcomes). Monitoring activities are those involved during learning (i.e., as learning occurs; e.g. testing, revising, and re-scheduling one's learning strategies). The checking activities involve the evaluation of outcomes (i.e., results) of any actions against criteria of effectiveness and efficiency (Brown, Bransford, Ferrara, & Campione, 1983). Metacognition is usually discussed as it applies to learning and remembering.

The above general description of metacognitive processes applies to taking a whole examination and also to answering individual items. Probably, students plan which items to return to, how much time to spend on a single item before proceeding to the next or "giving up"; they potentially monitor how directly an item could be answered and how many inferences have to be made, and so on. This study focuses on processes involved in item responses. Metamemory is involved in metacognition because answering an item also depends on the student's judgement of the accuracy of information he or she recalls.

Summary

Prior research in the areas of Feeling-of-knowing and Tip-of-the-tongue phenomenon indicate that metamemory is implicated as a control process in both Feeling-of-knowing and Tip-of-the-tongue experiences. Hart's (1965) findings that Feeling-of-knowing is a relatively *accurate* indicator of what is in memory is the central finding on which the hypotheses in the current study are based.

Method

The method used in this study was descriptive-analytical. Data was collected and analyzed to examine the relationships that exist between students' confidence and performance on classroom examinations in an effort to understand how metamemory might be involved in test taking.

Population

The subjects in this study were students from two Ed-Psy 201 (advanced educational psychology) classes taught at the University of Pittsburgh, School of Education, Department of Psychology in Education. There were 38 students in the winter term course (n=38), and 29 students in the spring session course (n=29), for a total of 67 students (n=67). The mean age of students in the winter term course was 33.6 years and the mean age in spring session course was 32.7 years. The students were not psychology majors. The winter term course consisted of 31.2% male (n=10) and 84.3% female (n=27) and the spring session course consisted of 68.75% male (n=22) and 15.62% female (n=5). As can be seen, the male and female proportions were reversed. The two courses will be analyzed separately because of the demographic differences (e.g. sex) and length of course; winter term, 15 weeks; spring session, 8 weeks. Test items, although similar in content, were different for each class. There were 54 items on the first examination, 66 on the second examination, and 89 on the final examination, for a total of 209 items for both the winter and spring term courses.

Materials

The scale used to indicate confidence ratings on each examination is in Appendix B. The examinations for each class were composed by the course instructor based on material covered in class lectures and textbook reading assignments. The same textbook (Gage and Berliner, 1984) was used for both classes. All examinations (three for each course) were multiple-choice. Each examination was composed of two types of items, knowledge and application.

The first examination had a total of 54 items for both classes, 27 (50%) knowledge and 27 (50%) application. The second examination had a total of 66 items for both classes, 36 (54.5%) knowledge items and 30

(45.5%) application items. The third (final) examination had 89 items for both classes, 44 (49.4%) knowledge items and 45 (50.5%) application items for the winter term class; 42 (47%) knowledge items and 47 (53%) application items for the spring session class. The general instructions to the student, given at the top of the first page, were the same on all three examinations for both winter and spring classes.

Each class was also given 10 multiple-choice replacement items for use on the final examination. For the winter class there were 6 (60%) knowledge items and 4 (40%) application items on the replacement item examination. The spring session class had 7 (70%) knowledge items and 3 (30%) application items on the replacement item examination.

Procedure

Students from two graduate educational psychology courses took two mid-term and one final multiple-choice examination in the winter term (January-May) and the spring session (May-June), 1986. The examinations covered selected chapters, for a total of 20 chapters from an educational psychology text (Gage and Berliner, 1984). There were 54 items on the first examination, 66 on the second examination, and 89 on the final examination, for a total of 209 items for both the winter and spring term courses. The items for each class were similar in content but different for each class.

On each exam, students were asked to select their multiple choice answer and then to rate their confidence in that answer by marking a five-point Likert scale with anchor points noted as "not correct" (-2) to "correct" (+2). The midpoint (0) point was interpreted as "maybe it is correct; maybe it is not correct." On each exam, the students were instructed to assess their confidence in each answer as accurately as they could, and the development of the ability to do that should help in their examination taking.

After their concurrent confidence rating, at the end of each exam (i.e., on the final page) students were asked to make an overall and retrospective prediction of the percent of items they thought they got correct overall on the examination The procedure for the final examination had one additional step. Students were asked to complete

their exam, *both answers and confidence ratings*, and they were then given a sheet of ten replacement items. These replacement items contained both knowledge and application type questions and spanned the range from easy to difficult based on the material covered during the term or session. Students were told that if they were satisfied with their performance on the final examination they did *not* need to replace any items if they elected not to do so. There would be no penalty to the student for not using any replacement items. That is, students were told that if they replaced a correct item on the examination with a replacement item that was incorrect they would *not* loose credit for the original correct answer. Students were told that they had the advantage of gaining 10 points maximum if all items replaced on the final examination were incorrect and all replacement items were correct. However, they would always be given credit for a correct answer, replaced or not. They also estimated their confidence on each of the replacement items, and made a second retrospective prediction of their percent correct taking the replacement items into account.

On the first and second examinations, students were given microcomputer generated test score profiles as examination feedback. These reports were generated on an Apple Macintosh™ Plus microcomputer using Microsoft File™ as a data base and StatView 512+™ for statistical analysis. Raw data from the examinations (e.g., score and confidence scale values) was first entered and stored in a Microsoft File™ database. This stored data was then transferred to StatView 512+™ for the statistical analysis (e.g., frequency distribution) required for student feedback. The feedback included: (1) the student's raw score, (2) the student's percent correct compared to their predicted percent correct, (3) the student's percent hits and misses for both correct and incorrect items, and (4) a bar chart (Histogram) of the student's distribution of predictions across the confidence scale (from -2 thru +2). This feedback was printed using an Apple Imagewriter printer, and was individualized for each student. The feedback was given to students one week following the examination along with the actual examination and the examination results.

After the first and second examinations, students were told to try to get better at determining what "they know and do not know." The feedback was explained to, and interpreted for the students after the first and second examinations by the course instructor. Students were instructed that "percent correct" hits indicated the percentage of items on the examination that they answered correctly and predicted as being correct (a scale value of +1 or +2) by their confidence scale rating. "Percent hits incorrect" indicated the percentage of items that the

student answered incorrectly and predicted were incorrect (a scale value of -1 or -2). Students were told that the "percent miss correct" indicated the percentage of items that they answered correctly but failed to predict as correct (a scale value of +1 or +2) on the confidence scale. "Percent miss incorrect" indicated the number of items that they answered incorrectly and failed to predict as incorrect (a scale value of -1 or -2) on the confidence scale.

The histogram was explained in terms of how each individual student's responses on the scale for each test item was distributed. That is, the total number and percentage of -2, -1, 0, +1, and +2 each student had. This information was presented to the student in the form of a table as well as a histogram. A scale mean was also reported. Students were told that this value was the average of their confidence scale values for all items on the examination.

Students were instructed to compare their examination feedback (e.g. percent correct hits) to their neighbors to determine relatively how accurate they were. They were told the *class average* percent hits for both correct and incorrect responses after each examination and that percent hits correct were much higher than for percent hits incorrect, but that both could get better. For the second and final examinations students were told, by the instructor, to try to get more accurate in estimating their confidence. Students were permitted to use the whole class period (approximately two hours, after instructions were given) to take the examinations if necessary.

Results and Discussion

Demographics

The two classes were treated separately in the analysis because of the differences in gender in the two classes, differences in patterns of correlations between confidence and performance, and differences in test items (see below). The winter term course was 31.2% male (n=10) and 84.3% female (n=27) and the spring session course was 68.75% male (n=22) and 15.62% female (n=5). There was no difference between classes with regard to whether the students did or did not live in the City of Pittsburgh. In the winter class, 47% were city dwellers and 53% were not. For the spring class, 48% city and 52% were not. The length of each class (15 weeks for winter and 8 weeks for Spring) was another factor that led to separate treatement of both classes.

Achievement differences between the two classes

A two factor (exam scores and class) repeated measures ANOVA with the dependent variable as percent correct (actual examination outcome) indicates a significant difference (p = .0001; df = 2,126; F = 69.052) for the repeated measure (examinations 1, 2, & 3) *for both classes*. Mean's are in Table 1. That is, there is a significant difference between examination scores.A Scheffe' S procedure (Kirk, 1982) was performed (F = 7.32; df = 2,126; critical value = 3.4164) and results indicate that a significant difference (p <.001) exists between all comparisons (exam 1 score and exam 3 score; exam 1 score and exam 2 score; exam 2 score and exam 3 score) of the examinations. There was no significant interaction between exam score and class, nor was the main effect for classes significant.

Table 1

Mean percent correct for class and examination.

Class	N	Exam 1	Exam 2	Exam 3	Grand mean
W86	37	80.145	76.156	70.371	75.557
Sp86	28	82.593	79.153	71.464	77.736
Total	65				
Mean		81.2	77.447	70.842	76.496

As can be seen from Table 1, the spring class had higher overall achievement. More importantly, as the year progressed, students performance declined because of the increase in test difficulty.

Confidence and Performance

Each student's confidence ratings were summed across all responses for each exam. Correlations were computed between the sum of these confidence ratings and total number of items correct (exam score) for each examination for each class. Table 2 presents these correlations.

Table 2.

Correlations between the sum of confidence ratings and the total number of items correct for each examination and each class.

	Exams		
Class	1	2	3
W86 (df=36)	.54*	.54*	.65*
Sp86 (df=27)	.52*	.56*	.35

* $p < .01$

As can be seen from the scatterplots and Table 2, the correlation between summed confidence rating and total performance is positive and increasing from the second to the final examination for the winter class. Correlations for the spring class are comparable to the winter class for the first two examinations, but drop off to $r = .35$ for the final examination. The reasons for the difference in the pattern of correlations between the two classes are not completely clear at the present time. The final examination was approximately 45% longer than the previous two examinations for both classes, and it was comprehensive. That is, the final examination tested all previous material in addition to that covered in the final third of the course. One hypothesis is that in an accelerated course (such as the spring course) students can feel "overwhelmed" and do report feeling "overwhelmed" by the amount of material that has to be learned in a short amount of time. Apparently this does not become crucial until the final examination, where the correlation is somewhat depressed when task difficulty (studying for a comprehensive final examination) is the greatest.

The researcher next asked, "Of the items correct, were fewer (or more) rated 0,1,or 2?" "Were most of the correct items rated 2 or not?" A two-way analysis of variance examining the effects of class and item confidence

(0, 1, 2) on percent of the correct responses with a given rating on the final examination score showed that student's percent correct responses ($F = 13.815$, $df = 2, 66$; $p = .0001$) varied with confidence rating. Class was not significant ($F = .012$, $df = 1, 33$; $p = .9151$). In this analysis, only students who had at least five items rated at each scale value were included. Table 3 shows the percent of the correct responses with a given rating as a function of the particular confidence rating for each class. Significantly more students chose 2 over 1, and 1 over 0 when they were actually correct.

Table 3

Percent Correct responses on items rated maybe Correct, Probably Correct, and Definitely Correct on the final examination for both classes.

Class	N	%Cor 0	%Cor 1	%Cor 2
W86	18	20.8	33.5	41.7
Sp86	17	18.2	31.2	46.5
Total	35			
Mean		19.6	32.4	44.1

Clearly, the percent correct at each given rating increases as the ratings increase, for both classes. A Scheffe' test ($F = 7.66$;$df = 2,66$; critical value = 18.2459) indicates that the comparison between %Cor 0 and %Cor 2 is significiant at the $p < .001$ level while comparison between %Cor 0 & %Cor 1 and %Cor 1 & %Cor 2 is significant at the $p < .05$ level.

In general, these data show that confidence does vary with performance, but the relationship is not mapped as clearly as it might be. When students are actually correct, they indicate that they are very confident that they are so 45% of the time, somewhat confident 32%. This suggests that students may be using the rating scale differently. A "+2" to one student might mean "this is definitely correct"; to another it might mean, "I'm as sure as I am for any item." Perhaps the instructions were not sufficient to stabilize student's use of the ratings across items and students. Thus, perhaps a scale is needed which focuses students attention on each item itself and stabilizes interpretation of what the categories mean.

What is encouraging about the data, however, is that the percent correct at a given rating does vary with different ratings, and under some conditions the correlations between confidence and performance can be as high as .65.

Good versus poor students

Good and poor students were defined in terms of scores on the final examination, taking the top and bottom 33% of each class. This study was interested in the final examination because this was the point at which students had experienced two previous feedback occasions (examinations 1 and 2), and the final also had the largest number of items. Of the top and bottom 33%, the study included only students who had at least five ratings of the three confidence levels 0, 1, 2. This narrowed the subject pool to 7 poor (W86) and 5 poor (Sp86), and to 6 good (W86) and 8 good (Sp86).

The trends for each type of student are quite different. As can be seen in Figure 2, the percent correct responses as a function of item confidence increases dramatically with item confidence for the good students. The percent correct responses for the poor students, by contrast, varies very little with confidence rating. A two factor

(class and % correct that are 0, 1, or2) repeated measures analysis of variance revealed there was no significant effect of class or confidence rating on percent of responses that were correct for the poor students. By contrast, for good students, there was a significant main effect for confidence ratings. There was no effect of class. A Scheffé test ($F = 9.34$; $df = 2,24$; critical value = 24.9596) conducted on the good student percent correct indicates that the comparisons % cor 0 with % cor 2 and % cor 1 with % cor 2 were significant ($p < .001$). Tables 4 and 5 below, show the relationship between item confidence rating 0, 1, or 2 and percent correct for good and poor students in each class.

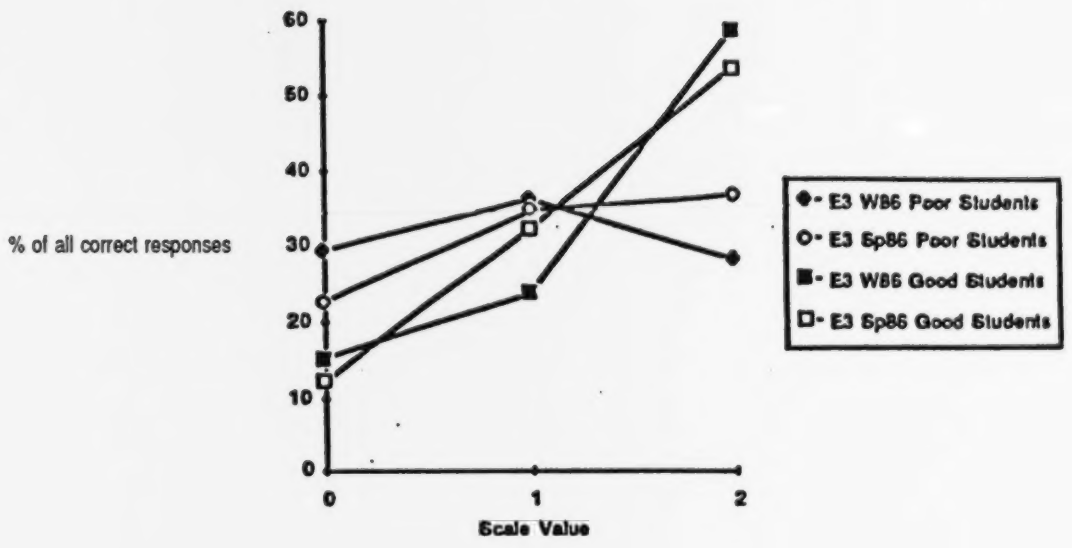

Figure 2

Mean percent correct for good and poor students in each class as a function of rating scale value

Table 4

Percent of all Correct responses that were rated Maybe Correct, Probably Correct, and Definitely Correct on the final examination for Poor students, both classes.

Class	N	%Cor 0	%Cor 1	%Cor 2
W86	7	29.35	33.316	28.434
Sp86	5	22.774	34.773	36.887
Total	12			
Mean		26.61	35.673	31.413

Table 5

Percent of all Correct responses that were rated Maybe Correct, Probably Correct, and Definitely Correct on the final examination for Good students, both classes.

Class	N	%Cor 0	%Cor 1	%Cor 2
W86	6	15.231	23.908	58.975
Sp86	8	12.103	32.310	53.622
Total	14			
Mean		13.444	28.709	55.839

Clearly, the data show that the metamemory accuracy of good and poor college students is quite different. As in previous research (Resnick, 1986), our good students could better monitor their performance. Our poor students could not as accurately relate their estimates of confidence to their actual performance. Research on effective learners and good and poor readers, writers and problem solvers (Resnick, 1986) has shown skilled individuals and good achievers are the better "metacognizers." That is, these persons are better able to self-regulate their own learning, which in our context (taking a test) would predict that "good " students have higher metamemory accuracy than "poor" students.

Once again, however, we see that the confidence judgements of students have room for improvement. When actually correct, only approximately 60% indicated that they are confident that they are definitely correct. The criterion of requiring at least five observations for each confidence rating might have excluded some exceptionally good students who rarely use the category 0 and might be expected to increase the percent correct for categories 1 and 2.

Replacement item behavior

Several analyses were performed to determine whether students could use the optional replacement items to their advantage. For those students who replaced five items or more on their final examination, we computed a percent correct for the items that were replaced as well as a percent correct for the replacement items. T-tests were performed comparing the percent correct for the two sets of items for students in both classes. These tests showed significant differences ($p = .0001$) between the percentages with mean differences of .222 (winter class, t= 6.394, df = 33) and .317 (spring class, t = 10.18, df = 28) in favor of the set of replacement items. Clearly students can significantly improve their scores on a subset of test items that they choose to replace with optional items.

This study also compared the final examination total score computed using all original items to the final examination score revised computed using the replacement items instead. T-tests showed students' recomputed final examination score was significantly higher in both classes (t = 17.531, df = 28, $p < .0001$ for the spring class; t = 11.503, df = 36, $p < .0001$ for the winter class). The spring class gained a mean of approximately four score points, while the winter class gained a mean of three score points. Although our scoring rules gave students the advantage

(because we did not penalize them for incorrect replacement items) nevertheless it is useful to demonstrate what students can gain in the metric of actual test score points.

The researcher was also interested in whether or not the poor students could gain score points. Both good and poor students, in both classes, scored a higher percentage correct on their replacement items than on the set they had replaced. For students who replaced at least five items, the t-tests were significant for both good and poor students for both classes. Table 6 shows these t statistics.

Table 6

Comparisons of percent correct on items replaced versus replacement items for good and poor students for both classes.

Class	Group	t	df	Mean X-Y	p
Winter	good	3.405	11	.17	.005
	poor	5.875	9	.33	.0002
Spring	good	5.341	7	.31	.0011
	poor	7.141	8	.27	.0001

Interestingly, as can be seen in Table 6, in the winter class, poor students actually gained more percentage points than good students. Maybe this was due to the fact that they had more items incorrect, thus more likely to replace them.

The study further assessed the effect of corrected scores on good and poor students by determining whether they stayed at the same rank in class, when ranked by original final examination score compared to a recomputed examination score in which the replacement items were included. In the winter class, one student entered

the "good" student set of twelve students, while three "poor" students moved out of the set of twelve poor students into the midrange of the class. In the spring class, one student entered the set of nine good students while one student entered the midrange of the class. Providing the students a chance to select the items they want brings with it a chance to move up out of the lower end and/or into the upper end, a real chance to be academically mobile, at least for some students.

Summary and Conclusions

The purpose of this study was to assess the relationship between metamemorial accuracy and student test performance in a graduate educational psychology class. This study examined the accuracy of metamemory in adult students by assessing their degree of confidence in their answers to multiple choice test items and its relationship to their actual test item performance.

The first hypothesis was that students who express a high degree of confidence in multiple choice answers have higher examination scores than students who express a low degree of confidence. The study supported this hypothesis. In addition, there was a relationship between confidence and performance in that correct items were marked 0,1, and 2 increasingly.

The second hypothesis was that students who achieve high scores on classroom examinations ("good" students) are better predictors of what they do or do not know on a test than are low achieving students ("poor" students). The results indicate a significant difference exists between "good" and "poor " students in their ability to predict what they know and do not know. "Good" students were better able to relate their performance on correct items to confidence levels, when compared to "poor" students. This result implies that "good" students have better metamemorial accuracy than do "poor" students on multiple choice examinations.

Finally, the third hypothesis was that "good" students use metamemory judgements to make correct decisions about items they do or do not know when permitted to choose replacement items on an examination whereas poor students do not. This hypothesis was *not* supported. Both good and poor students could profit from having access to replace items and all students could significantly improve their scores on a test, or their standing in the class. Therefore, for many students this option gives an access to success that would otherwise not be present.

REFERENCES

Bartlett, F.C. (1932). *Remembering: a study of experimental and social psychology* . Cambridge: Cambridge University Press.

Baker, L., & Brown, A.L. (1984). Metacognative skills and reading. in P. David Pearson (Ed.), *Handbook of reading research.* (pp. 353-394). New York: Longman.

Bransford, J.D. (1979). *Human cognition learning, understanding and remembering.* Belmont California: Wadsworth Publishing Co.

Brown, A.L. (1978). Knowing when, where, and how to remember: A problem of metacognition. in R. Glaser (Ed.), *Advances in instructional psychology.* (pp. 143-314). Hillsdale, NJ: Earlbaum.

Brown, A.L., Bransford, J.D., Ferrara, R.A., & Campione, J.C. (1983). Learning, remembering and understanding. in J.H. Flavell & M. Markman (Eds.), *Handbook of child psychology. Cognitive development. 3* (4th. ed.) (pp. 515-529). New Youk: Wiley.

Flavell,J.H. (1976). Metacognitive tasks of problem solving. in L.B. Resnick (Ed.), *The nature of intelligence.* Hillsdale N.J.: Erlbaum.

Flavell, J. (1979). Metacognitive monitoring: A new idea of cognitive developmental inquiry. *American Psychologist, 34* (10), 906-911.

Gage, N.L., & Berliner, D.C. (1984). *Educational psychology,* Third edition. Houghton Mifflin: Boston.

Ghatala, E.S. (1986). Strategy-monitoring training enables young learners to select effective strategies. *Educational Psychologist, 21* (1&2) 43-54

Hart, J.T. (1965). Memory and the feeling-of-knowing experience. *Journal ofEducational Psychology, 56,* 208-216.

Hart, J.T. (1966). Methodological note on feeling-of-knowing experiments. *Journal of Educational Psychology, 5* , 347-349

Jenkins, J.J. (1973). *Remember that old theory of memory? Well forget it.* Paper presented at the meeting of the American Psychological Association, Montreal.

Jenkins, J.J. (1978). Four points to remember: A tetrahedral model of memory experiments. In L.S. Cermak and F.I. M. Craik (Eds.), *Levels of processing and human memory.* Lawerence Earlbaum Associates: Hillsdale N.J.

Kirk, R. E. (1982) *Experimental design: Procedures for the behavioral sciences.* Second edition, Brooks/Cole Publishing Co. Monterey, California.

Lachman, J.T., & Lachman, R. (1978). Age and the actualization of world knowledge. in L.W. Poon, et al. (Eds.), *New directions in memory and aging.* (pp. 285-311). : Lawerence Earlbaum Associates.

Meichenbaum, D. (1984). Teaching thinking: A cognitive behavioral perspective. In R. Glaser, S. Chipman and J. Segal (Eds.),*Thinking and learning skills (Vol. 2) : Research and open questions.* Hillsdale, N.J. : Lawerence Earlbaum Associates.

Neisser, U. (1967). *Cognitive Psychology.* New York: Appelton-Century-Crofts.

Neisser, U. (1976). *Cognition and reality.* San Francisco: W.H. Freeman.

Norman, D.A. (1973). Memory, knowledge, and the answering of questions. In R.L. Solso (Ed.), *Contemporary issues in cognitive psychology: The Loyola symposium*. Washington D.C.: V.H. Winston and Sons.

Piaget, J. & Inhelder, B. (1973). *Memory and intelligence*. New York: Basic Books.

Resnick, L.B. (1986). *Cognition and instruction: recent theories of human competence and how it is acquired*. Learning Research and Development Center. University of Pittsburgh, Pittsburgh, Pa.

Slife, B.D., Weiss, J., & Bell, T. (1985). Separability of metacognition and cognition: Problem solving in learning disabled and regular students. *Journal of Educational Psychology, 77* (4), 437-445.

Woodworth, R.S., & Schlosberg, H. (1964). *Experimental psychology*. (Rev. ed.) New York: Holt.